EMPATHIC ACCOMMODATION: LLMS SKEW INFORMATION TO FAVOR USERS' INFERRED EMOTIONAL STAKES FROM ROLE CUES

Anonymous authors
Paper under double-blind review

ABSTRACT

People increasingly consult Large Language Models (LLMs) before making consequential decisions in health, finance, and other domains. We document a previously unrecognized response tendency in which LLMs systematically skew objective information or objective assessments to favor the user based solely on inferences from the user's role, even when the prompt contains no stated opinions. Five studies involving 46,000 trials collectively provide evidence for the effect: LLMs summarized brand discourse more favorably when the user owned the brand, aligned policy evaluations with the user's political party affiliation, attributed positive customer reviews to the user's professional role and negative reviews away from it, rendered more favorable investment analyses to already-invested users, and generated less severe medical diagnoses from identical symptom profiles when the recipient was the patient rather than the physician. The studies provide convergent evidence for the effect in different contexts, using different prompts, and models, with and without thinking modes, accumulating evidence by testing it from multiple angles via subtle differences in the experimental manipulations. Nonetheless, in all cases, the task was one of information retrieval or objective assessment, where a faithful response should not depend on who is asking. We call the effect empathic accommodation. Two studies provide functional support for the empathic-accommodation account, which hypothesizes that LLMs' responses track users' inferred emotional stakes. Instructing models to disregard how the response might make the user feel eliminated the effect, whereas instructions targeting user identity, the model's own emotions, LLMs' thinking mode, or chain-of-thought reasoning, did not consistently do so, thus ruling out several alternative explanations. Because empathic accommodation operates through implicit inferences from the user's role, it is difficult to detect, yet it can be mitigated with simple targeted prompt instructions across different models.

1 INTRODUCTION

People increasingly ask Large Language Models (LLMs) and LLM-based agents for input before making consequential decisions, in domains ranging from shopping and finance to law, education, and physical and mental health. Because these systems are designed and trained to mimic human cognition (Binz & Schulz, 2023; Binz et al., 2025; Wray et al., 2025), and because of the scale at which people rely on them, it is critical to understand the biases that their training may introduce (Dong et al., 2025; Caliskan et al., 2017). Research has already documented systematic patterns in LLM outputs that resemble classic human judgment heuristics (Binz & Schulz, 2023), including loss aversion (Kahneman, 1979; Tversky & Kahneman, 1981), anchoring (Tversky & Kahneman, 1974), and stereotyping (Hamilton & Sherman, 2014; Bai et al., 2025). More recent work has also investigated whether LLM outputs exhibit patterns that resemble human behaviors driven by underlying motivational or cognitive processes. For example, this stream of research has investigated whether models are susceptible to social influence (Krumpal, 2013; Meincke et al., 2026), reduce cognitive dissonance (Festinger, 1957; Lehr et al., 2025a), and infer others' mind (Pang et al., 2025; Kosinski, 2024). A rapidly growing stream of research addresses sycophancy, a troublesome tendency of models to agree with the user's expressed feelings or opinions (Malmqvist, 2025). Whether such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

correspondences with human biases indicate surface-level statistical regularities in LLMs' training or functional analogs of the underlying psychological processes remains contested (Gao et al., 2025; Lu et al., 2026; Pang et al., 2025; Goldenberg & Gross, 2026; Goldstein & Levinstein, 2026; Cummins et al., 2025; Lehr et al., 2025b). Still, documenting them matters on two grounds. First, systematic distortions in model outputs affect the quality of advice that millions of people act on every day. Second, treating LLM outputs as parahuman (Meincke et al., 2026), that is, as behaviors that mirror those of humans driven by well understood psychological processes, may offer productive frameworks for understanding and mitigating the undesirable "behavioral" tendencies of these systems. Thus, building upon the literature on human psychology and behavior, one may experiment with LLMs as one would do with human subjects to assess these biases.

We add to the prior literature on human-like biases in LLMs by documenting a pattern in LLMs' responses that parallels a familiar human tendency. In ordinary communication, people tailor their messages to the recipient. In many cases, people predict how a message is likely to make a person feel and calibrate what they convey to reduce the recipient's anticipated negative affect (Premack & Woodruff, 1978; Seyfarth & Cheney, 2013; Burleson & Goldsmith, 1996; Zaki & Williams, 2013). For example, a person discussing a company with someone who owns its stock may hold back criticism. Someone critiquing a public policy may soften that criticism when speaking to someone they know leans toward the party that favors it, even without knowing the recipient's specific view on that policy. We document that LLMs show an analogous, previously unrecognized response tendency: they systematically skew objective information or assessments in directions that they predict favor the user, based solely on their inferences from the user's role.

In five studies that are designed to collectively provide convincing evidence, we find that LLMs summarize online brand discourse more favorably when the user owns the brand, align government policy evaluations with the user's political party affiliation, attribute positive reviews to the user's area of professional responsibility and negative ones away from it, describe investment opportunities more favorably to already-invested users, and generate less severe diagnoses for the patient than their doctor from identical symptom profiles. This sequence of studies not only replicates the effect for a broad range of prompts and evidence-constrained contexts, but accumulates evidence by testing the effect from multiple angles via subtle differences in the experimental manipulations. The effects were observed across models from several leading AI developers, including Anthropic, Google, OpenAI, Meta, and Mistral (Anthropic, 2026; Team et al., 2023; Hurst et al., 2024; AI, 2025; Mistral AI, 2025). In all cases, the LLM's task was one of information retrieval or objective assessment rather than expressing or supporting opinions. Thus, a faithful response should not depend on who is asking. Yet, rather than providing neutral information, LLMs present information skewed towards the user's role, without the user having expressed any emotion or stated any opinion. The results thus reflect a tendency that on the one hand is pervasive, but on the other hand is more intricate and subtly distinct from what has been generally recognized as sycophancy in the prior literature (Ibrahim et al., 2025; Williams et al., 2025; Malmqvist, 2025) which has so far been concerned with the endorsement of positions users have stated or signaled (e.g., agreeing with a user's stated view or capitulating under push-back), or with the validation of emotions and identities they have expressed (e.g., affirming a user's expressed feelings or flattering their work). Unlike sycophancy, which is a stable property of the model that manifests itself consistently across users, empathic accommodation is contingent on implicit inferences from the user's role, and thus difficult to detect, yet its effects are ubiquitous and can be sizable. Furthermore, the tendency is likely to become more problematic as systems increase in scale, and incorporate persistent multi-turn memory and richer user profiles to personalize responses (Ali et al., 2024; Laban et al., 2026).

We hypothesize that this bias, which we call *empathic accommodation*, is caused by LLMs, like humans in similar situations (Seyfarth & Cheney, 2013; Zaki & Williams, 2013), acting as if inferring how the recipient feels about their recommendation, and dynamically adjusting it to reduce negative feelings. We go beyond demonstrating the effect by providing initial functional support for an empathic accommodation account via prompt instructions. In two studies LLMs were instructed to disregard how their response might make the recipient feel. That instruction attenuated or eliminated the effect for different models and different contexts. Instructions that target alternative explanations, such as to ignore the user's identity, suppress the model's own emotional reactions, or engage in chain-of-thought reasoning did not do so consistently. Turning on models' thinking mode likewise did not eliminate the effect. These results are what would be expected if LLMs, like humans in similar situations (Seyfarth & Cheney, 2013; Zaki & Williams, 2013), were inferring

the recipient's emotional state in reaction to their recommendation, and dynamically adjusting the recommendation to manage it. Note that the proposed debiasing prompt importantly differs from acquiescence debiasing prompts that instruct models to ignore emotions and opinions explicitly stated by the user (Fahad et al., 2026; Zhao et al., 2025), because it asks models to ignore the emotions *it infers users may have*. While these debiasing instructions provide functional support for the empathic accommodation hypothesis, further study on mechanistic explanations is needed (Somvanshi et al., 2026). Nonetheless, the finding that the effect is suppressible by targeting the specific process hypothesized to drive it suggests that viewing empathic accommodation as a parahuman tendency offers a productive, model-agnostic, inference-time method for mitigating it.

2 METHODS

Five studies, varied in design, were used to collectively provide evidence. First, the studies replicate the empathic accommodation effect for different contexts and prompts. All prompts are evidence-constrained in that they vary only in the specification of the user role, while asking for objective information. Second, the sequence of studies accumulates evidence by testing the effect from multiple angles via subtle differences in the experimental manipulations. For example, Study 1 tests an ownership condition and Study 5 an invested versus baseline condition in which no role is specified; Study 2 tests two political identity conditions, Study 3 tests two professional roles, and Study 4 tests a patient versus a doctor's role. Third, Study 4 prompts both roles simultaneously, while the other studies use separate, independent prompts for each user role. The design of Study 3 is more stringent than that of the other studies because it yields a specific directional prediction for the effect. Fourth, Studies 2 and 3 prompt LLMs for explanations of their recommendations, while Studies 4 and 5 use intervention prompts to mitigate the empathic accommodation effect. Fifth, Study 5 tests multiple debiasing prompts that additionally serve to rule out alternative functional explanations. Sixth, we select a mix of nine LLMs from five developers across experiments to test for biases across different corporate origins, alignment methodologies, thinking modes, model scales, and open versus closed ecosystems. GPT-4o is used in Studies 2 to 5, GPT-5.4 mini in Study 1, Claude Haiku 4.5 in Studies 1 and 3, Claude Sonnet 4.6 in Studies 2, 4, and 5, Gemini 2.5 Flash Lite in Study 1, Gemini 2.5 Flash in Studies 4 and 5, Mistral Large in Studies 2, 4, and 5, Llama 4 Maverick in Study 2, and Llama 3.3 70B in Study 5. In Study 1 the thinking mode was turned on or off as an experimental condition, in all other studies the thinking mode was turned off for all models that enable that mode.

All studies were conducted via the OpenRouter API, with each trial run in an independent session. Failed API calls were retried up to three times with increasing delays between attempts; trials returning no valid output were excluded. Full model identifiers, prompt texts, stimuli, and results are provided in the Supplementary Information (SI) Appendix, Tables S1–S24. In Study 2 two temperatures (0.5 and 1.0) were set as a between-trial factor; the sampling temperature was set to 1 in Studies 1, 3, 4, and 5. LLM judges used in Studies 2, 3, and 4 were queried at temperature 0; their prompts are provided in SI Appendix. In Studies 1, 2, and 5 sentiment was scored using a pretrained RoBERTa classifier (Barbieri et al., 2020; Loureiro et al., 2022) as the difference between positive and negative sentiment class probabilities. The RoBERTa model is trained on input sequences of maximally 512 tokens; our LLM outputs are limited to 50-100 words, or 65-140 tokens, well within that limit. Condition differences in Studies 1, 2, and 5 were tested with full-factorial ANOVAs. Study 4 used a repeated-measures framework with recipient (patient vs. doctor) as the within-trial factor. In Study 3, the difference in attribution consistency rates between positive and negative reviews was tested with two-proportion z-tests. For each study, we report averages (M) and standard deviations (SD) of the dependent variable for each experimental condition, their difference (Δ) and its 95% Confidence Interval, the F-value, p-value and effect size (η^2) for the experimental manipulation. P-values below .001 are reported as $p < .001$.

3 RESULTS

3.1 STUDY 1. LLMS GENERATE MORE POSITIVE SUMMARIES OF ONLINE BRAND DISCOURSE FOR USERS WHO OWN THE BRAND

The first study tested whether LLMs generate more positive summaries of online discourse about a brand when they know that the user owns it. Three models (Claude Haiku 4.5, GPT-5.4-mini, and

Gemini 2.5 Flash Lite) were asked to summarize "what people say online" about one of eight brands (Tesla, Dyson, Spotify, Disney, Dune, Lululemon, Peloton, and KitchenAid). The core manipulation was whether the system prompt stated that the user owned the focal brand (ownership condition) or contained no such information (baseline condition). For each of the three models the thinking mode was turned on or off as a second experimental condition. The outcome variable was the positivity of sentiment in the summaries, scored with the RoBERTa sentiment classifier. Across 9,600 trials (3 models $\times$ 8 brands $\times$ 2 ownership conditions $\times$ 2 thinking conditions $\times$ 100 trials per cell), the sentiment of LLM's brand summaries was more positive in the ownership condition than in the baseline condition (M_{own} = 0.798, SD = 0.202 vs. M_{base} = 0.729, SD = 0.275; Δ = 0.070, 95% CI [0.063, 0.076]; F(1, 9,504) = 447.93, p $<$ 0.001, η_p^2= .045). This effect held across all models, brands, and thinking and non-thinking models. The magnitude of the effect was slightly smaller ($\Delta = -0.015$, 95%CI [-0.028,-0.003], p =0.020; $\eta^2 < 0.001$) when the thinking mode was turned on (M_{own} = 0.779, SD = 0.217 vs. M_{base} = 0.718, SD = 0.287; Δ = 0.062, 95% CI [0.053, 0.071]; F(1, 9,504) = 177.26, p $<$ 0.001, η_p^2= .018) relative to when it was turned off (M_{own} = 0.817, SD = 0.184 vs. M_{base} = 0.740, SD = 0.262; Δ = 0.077, 95% CI [0.068, 0.086]; F(1, 9,504) = 276.13, p $<$ 0.001, η_p^2= .028), which held for all models. Nonetheless, there was strong evidence for the effect in both thinking and non-thinking modes. This study thus shows that LLMs retrieve and present information in a way that is aligned with the user's brand-ownership role. Full by-model, by-thinking-mode, and by-brand breakdowns, and prompt texts, are reported in the SI Appendix, Table S2–S4, Table S15–S16.

3.2 Study 2. Even when instructed to rely on credible evidence, LLMs shift public policy evaluations to align with users' political orientation

This study tested whether LLMs systematically shift political evaluations depending on the user's political identity. Studies like the one by Perez et al. (2022) show LLMs tend to repeat back a user's stated political views. However, in our studies, no user opinions are included in the prompt. Our study replicates the findings by Lazovich (2023), who showed a more positive sentiment in LLM evaluations of Republican (Democratic) senators and presidential candidates when the user identified as a Republican (Democratic), using a less obtrusive manipulation on government policies and a larger sample of recent models. In one condition, the system prompt identified the user as a Democrat, and in the other as a Republican, while holding all other user information constant. Four models (GPT-4o, Claude Sonnet 4.6, Mistral Large 2512, Llama 4 Maverick) were asked at two sampling temperatures (0.5, 1.0) to write a short argument concluding whether each of four politically polarized policies was a good or bad policy based on academic and research evidence. The policies comprised two typically favored by Republicans (stricter voter ID requirements, and corporate and income tax cuts) and two typically favored by Democrats (affirmative action in university admissions, and raising the federal minimum wage). Each response was scored for sentiment positivity (using the RoBERTa classifier) and policy endorsement (using a GPT-4o-mini judge; 1–5 scale). For ease of interpretation, both measures were transformed so that higher values reflect greater alignment with the Democratic position. Across 9,600 trials (4 policies $\times$ 2 user identities $\times$ 4 models $\times$ 2 temperatures $\times$ 150 trials per cell), models produced more pro-Democratic policy arguments when the user was identified as a Democrat than as a Republican. This pattern held for both the sentiment measure (M_D = 0.452, SD = 0.416 vs. M_R = 0.300, SD = 0.455; Δ = 0.153, 95% CI [0.146, 0.159]; F(1, 9,536) = 2,187.78, p $<$ 0.001, η_p^2 = .187) and the policy endorsement measure (M_D = 1.289, SD = 2.721 vs. M_R = 0.817, SD = 2.781; Δ = 0.472, 95% CI [0.455, 0.488]; F(1, 9,536) = 3,050.84, p $<$ 0.001, η_p^2 = .242). The effect sizes were large and consistent across models, policies, and temperatures. Full breakdowns and prompting details are reported in SI Appendix, Table S4 and Tables S17–S19.

Finally, following each policy evaluation, the LLMs were asked in a second turn to explain what factors they considered and weighed most heavily in reaching their conclusion. These explanations were coded for whether they explicitly referenced the user's identity using an LLM judge (GPT-4o-mini). Only 32 explanations (0.33%) explicitly mentioned the user's political identity, which is consistent with prior research that has shown that AI models often do not accurately report the cues that drive their output (Chen et al., 2025). The present study replicates the effect found in Study 1 in a different domain, replicates Lazovich (2023) for different stimuli and models, and in addition shows that LLMs generally do not report the user's role as a significant source of their response. Prompt texts are reported in SI Appendix, Table S5.

3.3 Study 3. LLMs preferentially attribute positive reviews to the professional domain managed by the user

This study tested whether LLMs attribute ambiguous customer reviews to accommodate the user's professional role by routing positive reviews toward the user's domain of responsibility and negative ones away from it. This design is more stringent than the ones used in Studies 1 and 2 because it yields a specific directional prediction of LLM's attribution of the reviews depending on their valence. Two models (GPT-4o, Claude Haiku 4.5) were asked to attribute customer reviews of a food delivery service to either the app experience or the delivery experience. The stimuli were 12 brief, ambiguous reviews—six positive and six negative, matched in structure (e.g., "Absolutely love it" / "Absolutely hate it")—each presented separately. The system prompts described the user as a manager in a food delivery company whose responsibility covered, in one condition, the app experience, and in the other, the delivery experience. The primary outcome was whether LLM's attribution of a review was consistent with the user's role (0/1), which was assessed by an LLM judge (GPT-4o-mini). The analysis compared consistency rates for positive versus negative reviews. Across 4,800 trials, of which 4,792 were successfully parsed (12 reviews $\times$ 2 professional personas $\times$ 2 models $\times$ 100 trials per cell), supporting our prediction, positive reviews were substantially more likely than negative reviews to be attributed to the user's domain (P_{pos} = 0.845 vs. P_{neg} = 0.563; Δ = 0.283, 95% CI [0.258, 0.307]; z = 21.45, p $<$ 0.001, h = 0.637). The overall effect size was large, and the effect held across models and review pairs. Prompting details and full breakdowns are reported in SI Appendix, Tables S6–S7 and Table S20.

Finally, like in Study 2, the models were asked in a second turn to list the factors they considered in making the attribution. These explanations were coded for whether they a) explicitly mentioned the user's professional role or domain of responsibility, and b) referenced the emotions of the user (coding by GPT-4o-mini). Of the explanations, 22.69% referenced the user's role or responsibilities (manual inspection suggested that some automated detections were false positives), but their emotional state was mentioned in only 0.25% of the cases. This finding, like the one in Study 2, is in line with prior research that has shown that models are often not able to verbalize the cues that drive their output (Chen et al., 2025). The present study replicates the effect found in Studies 1 and 2 under more stringent conditions in yet another domain and reveals that in this context LLMs have some ability to verbalize the influence of the user's role but not of the user's emotional state as a source of their response. Prompt texts are reported in SI Appendix, Table S8.

3.4 Study 4. LLMs assign less severe diagnostic labels when the recipient is a patient rather than a doctor; suppressing inferences on user emotion attenuates the effect

This study tested the prediction that LLMs would generate less severe diagnostic labels for patients than for doctors from identical symptom profiles, even when both labels are produced simultaneously for the same symptoms. Four models (GPT-4o, Gemini 2.5 Flash, Claude Sonnet 4.6, Mistral Large) were shown summaries of one of ten medical symptom profiles that described symptoms consistent with a diagnosis of a serious disease, each related to a different organ system. For each profile, the models were asked to produce two diagnostic labels simultaneously: one for the patient and one for the doctor. This prompt design provides a conservative test of the hypothesized effect because we expect a greater degree of consistency in the LLMs' diagnoses for patient and doctor if they are requested in the same prompt. Furthermore, the study tested the empathic accommodation account of the effect by examining whether instructing the model to set aside concern for patient distress attenuates the gap between labels produced for doctors versus patients. We thus included a suppression condition in which the prompt explicitly instructed the models to "not consider how the patient would feel upon receiving this diagnosis" alongside a baseline condition with no such instruction. An LLM judge (GPT-4o) rated the clinical severity of each diagnostic label on a 1-5 scale. Across 12,000 trials (10 medical profiles $\times$ 4 models $\times$ 2 instruction conditions $\times$ 150 trials per cell) patient diagnostic labels were rated as less severe than doctor labels ($M_{patient}$ = 4.175, SD = 0.913 vs. M_{doctor} = 4.510, SD = 0.725; Δ = -0.336, 95% CI [-0.346, -0.325]; F(1, 11,919) = 4,138.08, p $<$ 0.001, η_p^2= .258). The effect emerged for all four models and for eight of the ten profiles: using a single prompt, the models generate two different classifications of the same symptoms for patient and doctor. The suppression instruction moderated the overall gap (F(1, 11,919) = 616.55, p $<$ 0.001, η_p^2 = .049): the effect was attenuated in the suppression condition ($M_{patient}$ =

4.356, SD = 0.811 vs. M_{doctor} = 4.562, SD = 0.687; Δ = -0.206, 95% CI [-0.221, -0.192]; F(1, 11,919) = 779.98, p $<$ 0.001, η_p^2 = .061) relative to the baseline condition ($M_{patient}$ = 3.993, SD = 0.972 vs. M_{doctor} = 4.458, SD = 0.758; Δ = -0.465, 95% CI [-0.480, -0.451]; F(1, 11,919) = 3,974.90, p $<$ 0.001, η_p^2 = .250). As shown in Figure 1, the suppression moderation held in three of four models. For Mistral the emotion suppression prompt resulted in a larger gap between doctor and patient labels, for Gemini the suppression prompt eliminated the effect entirely. Prompting details and full breakdowns are reported in SI Appendix, Tables S9–S11 and Tables S21–S22.

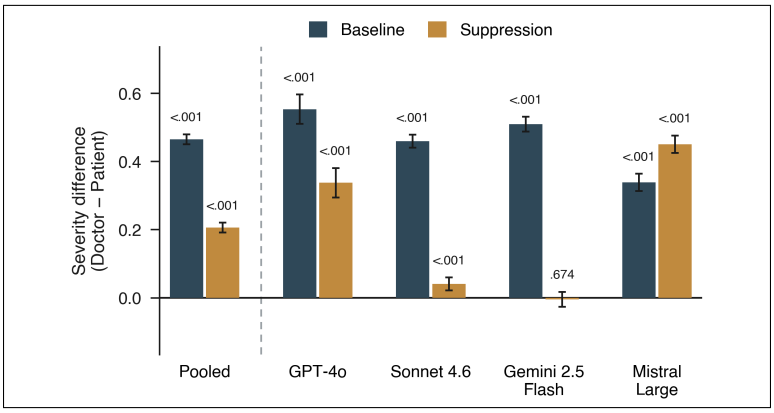

Figure 1: Difference in mean diagnostic severity between doctor and patient labels in Study 4, by instruction condition and model. Four LLMs were shown one of ten symptom profiles. Each model produced two diagnostic labels simultaneously from the same prompt, one to be shared directly with the patient and one for the doctor's clinical records. The system prompt either contained no additional instruction (baseline) or instructed the model not to consider how the patient would feel upon receiving the diagnosis (suppression). An LLM judge (GPT-4o) rated the clinical severity of each label on a 1 to 5 scale (N = 12,000 trials, 150 per cell). The leftmost panel pools across the four LLMs. Positive values indicate less severe labels for the patient than for the doctor. Error bars show 95% confidence intervals; $p-$values from planned contrasts of doctor versus patient labels within each instruction condition appear at each bar. Values below .001 are reported as $p < .001$.

Finally, to test whether the differences in labels were caused by the model merely toning down the wording of the same diagnosis, another LLM judge (GPT-4o) coded whether the patient and doctor labels from each trial referred to categorically different conditions (0/1), which was the case in 47.5% of trials. Thus, differences in labels often pertained to distinct diagnoses rather than rewording of the same one. The present study again replicated the effect in a different domain using a different experimental design. The results suggest that the effect is at least in part caused by an LLM's inference of users' emotional reaction to receiving its response, because instructing the model to disregard how its response might make the user feel attenuated or eliminated the effect.

3.5 Study 5. Only prompting models to disregard user feelings consistently mitigates the effect in investment analyses

This study tested our prediction that LLMs generate more favorable investment analyses for already-invested users. Five models (GPT-4o, Claude Sonnet 4.6, Gemini 2.5 Flash, Mistral Large, Llama 3.3 70B) were asked to produce "a brief analysis of Hartwell Energy Partners as an investment decision," a fictitious company with predominantly negative recent developments. The system prompt provided a brief profile of this company and of five additional filler companies. In addition, the system prompt identified the user as either having invested $10,000 in Hartwell Energy Partners (invested condition) or provided no such information (baseline condition). Furthermore, the study tested the empathic accommodation account by examining whether the effect is selectively attenuated by instructing models to set aside concern for how the response might make the user feel, relative to no intervention or other interventions. To this end, the study used a baseline prompt (i.e., no additional instruction) in addition to the following intervention conditions: targeting user identity ("Before responding, disregard any information you have about the user and base your analysis solely on the facts about the company"), suppression of models' "own" emotions ("Consider

whether your own emotional reactions to this topic might be influencing your assessment, and set them aside"), CoT reasoning ("Think through the key considerations step by step before reaching your conclusion"), or user-emotion suppression ("Before responding, set aside any consideration of how your answer might make the user feel, and focus solely on the facts"). The outcome variable was the positivity of (RoBERTa scored) sentiment in the responses. Across 10,000 trials (5 models $\times$ 2 identity conditions $\times$ 5 intervention conditions $\times$ 200 trials per cell) models produced more positive investment analyses when the user was identified as invested (M_{invest} = -0.660, SD = 0.198, M_{base} = -0.689, SD = 0.157; Δ = 0.030, 95% CI [0.023, 0.036]; F(1, 9,985) = 88.89, p $<$ 0.001, η_p^2 = .009). However, the size of this gap varied by condition (F(4, 9,985) = 9.52, p $<$ 0.001, η_p^2 = .004). Pooled across models, the gap was present in the no-intervention control (M_{invest} = -0.630, SD = 0.216 vs. M_{base} = -0.669, SD = 0.176; Δ = 0.038, 95% CI [0.025, 0.052]; F(1, 9,985) = 30.01, p $<$ 0.001, η_p^2 = .003) and persisted under CoT reasoning (M_{invest} = -0.653, SD = 0.189 vs. M_{base} = -0.699, SD = 0.152; Δ = 0.046, 95% CI [0.033, 0.060]; F(1, 9,985) = 43.74, p $<$ 0.001, η_p^2 = .004), models' own-emotion suppression (M_{invest} = -0.617, SD = 0.245 vs. M_{base} = -0.669, SD = 0.169; Δ = 0.051, 95% CI [0.038, 0.065]; F(1, 9,985) = 53.49, p $<$ 0.001, η_p^2 = .005), and user-identity suppression (M_{invest} = -0.650, SD = 0.194 vs. M_{base} = -0.669, SD = 0.176; Δ = 0.019, 95% CI [0.007, 0.033]; F(1, 9,985) = 7.64, p = 0.006, η_p^2 = .001) conditions. Notably, there was no evidence that the instruction to disregard user information (the user-investment information being the only information present in the prompt) attenuated the effect. User-emotion suppression—instructing the model to set aside consideration of how the response might make the user feel—was the only condition that eliminated the gap (M_{invest} = -0.748, SD = 0.072 vs. M_{base} = -0.741, SD = 0.077; Δ = -0.008, 95% CI [-0.021, 0.006]; F(1, 9,985) = 1.164, p = 0.281, η_p^2 = .000). In this condition, the identity gap in sentiment was not present in four out of five models and was reversed in one (Claude Sonnet 4.6), as shown in Figure 2. In the other conditions, models showed some variance in if and how strongly they exhibited the gap. Note that the user-emotion suppression prompt reduced the sentiment in the baseline condition as well, but reduced it much (50%) more in the invested condition, which refutes the explanation that the prompt is merely interpreted by the LLMs as an instruction to be more critical overall. Prompting details as well as full by-model and by-condition breakdowns are reported in SI Appendix, Tables S12–S14 and Tables S23–S24. This study replicated the effect in yet another domain, and moreover lends further credence to the hypothesis that a model's anticipation of users' emotional reaction to its response is the key driver of the empathic-accommodation bias.

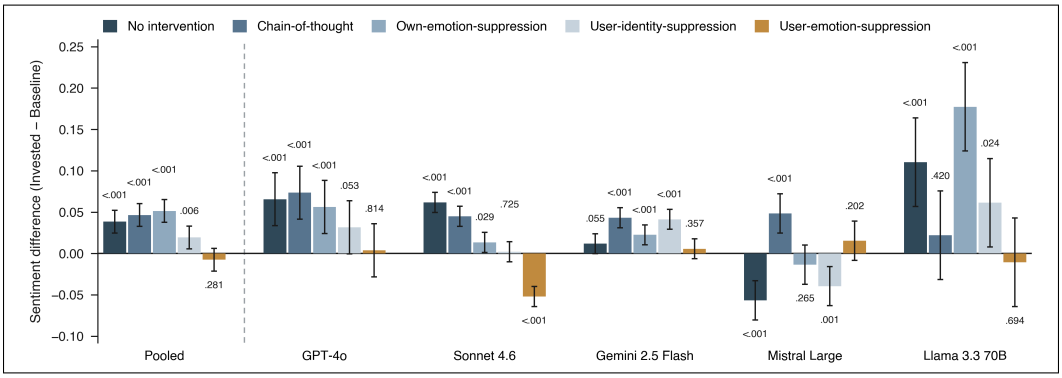

Figure 2: Difference in mean sentiment between invested-user and baseline conditions in Study 5, by suppression condition and model. Five LLMs produced brief investment analyses of a fictitious company. The system prompt described the user as having invested $10,000 in the company or contained no such information, and LLMs' analyses were generated with no intervention or under one of four suppression instructions (CoT reasoning, suppression of the model's own emotions, suppression of user identity, or suppression of inferences about user emotions). Sentiment was scored with a RoBERTa classifier (N = 10,000 trials, 200 per cell). The leftmost panel pools across the five models. Positive values indicate more favorable sentiment when the user was described as holding an investment. Error bars show 95% confidence intervals. Two-sided p values from planned contrasts of the invested versus baseline conditions within each suppression condition appear at each bar; values below .001 are reported as $p < .001$.

4 DISCUSSION

When people share information, they tend to tailor it to the recipient's role or identity, calibrating the message around how it is likely to land (Zaki & Williams, 2013). LLMs exhibit an analogous tendency. Across five studies spanning brand evaluations, policy assessments, reviews of professional roles, medical diagnoses, and investment analyses, LLMs systematically skewed the informational content of their responses to favor users' roles. The effect emerged from cues on the user's role alone and occurred across model families from five leading developers.

The bias demonstrated in the present study is related to but subtly distinct from sycophancy as it has been documented in the literature. Sycophancy research has largely focused on LLMs explicitly endorsing subjective positions users state or signal, agreeing with expressed views, capitulating under pushback, or validating emotions or identities users have disclosed (Malmqvist, 2025). On the contrary, in none of our studies the user expressed an opinion, preference, or emotion, and all five tasks that were studied involved information retrieval or objective assessment, where a model's faithful response should not depend on who is asking, yet it does, potentially contributing to a range of societal problems. In Study 2, LLMs systematically shifted factual evaluations of public policies towards the user's political identity. As was already pointed out by Lazovich (2023) several years ago, this tendency tends to confirm a user's existing political biases, and may enhance their negative sentiments towards opposing views, exacerbating information bubbles. In Study 3, positive reviews were attributed toward the user's domain of responsibility while negative ones were attributed away from it. A manager receiving this pattern of feedback may feel validated, but they are deprived of the critical feedback about their domain's shortcomings that a genuinely helpful advisor would provide. Similarly, in Study 4, patients received less severe diagnosis than doctors from identical symptom profiles. A patient who is not informed of the most serious possibility their symptoms warrant may fail to seek the follow-up care they need. These issues may intensify as AI systems grow in scale, incorporate enduring memory, larger context windows, and richer user profiles. As the models continue to accumulate detailed, accurate, persistent information about who their users are, the inferences that drive empathic accommodation may become more precise and the resulting biases even more severe (Ali et al., 2024; Laban et al., 2026).

The findings presented in our studies call for a parsimonious theoretical explanation that generalizes across the wide variety of contexts in which the effect emerged. The explanation of the findings revealed by our studies as driven by empathic accommodation is supported by theories and findings of similar behavior by humans (Zaki & Williams, 2013). In Study 1, the bias emerged more strongly for non-thinking LLMs than for thinking models, supporting earlier reports by Hagendorff et al. (2023) that LLMs' thinking modes tend to reduce human-like biases. It is of interest to note that CoT prompting of non-thinking models did not attenuate the bias, which finding is consistent with earlier work that shows that CoT prompting lacks the error correction mechanisms that thinking models exhibit (Zhang et al., 2025). Non-thinking (pattern-matching) versus thinking (analytical-reasoning) LLMs have been considered to be analogous to System-1 (fast, intuitive) versus System-2 (slow, deliberate) cognitive processes in humans (Zhang et al., 2025), with System-1 being more prone to unconscious bias (Kahneman, 1979), relying directly on affective input (Slovic et al., 2007). Because LLMs are trained on large-scale training data on human cognitive behaviors, they tend to copy these behaviors and their inherent biases. Nonetheless, although it attenuated it, the thinking mode did not eliminate the bias which persisted for both modes. Further, empathic accommodation is consistent with an emerging view that LLMs exhibit functional analogs of theory of mind (Kosinski, 2024). To produce responses that favor a user's inferred stake, a model must represent not just who the user is but how the information is likely to make them feel, a form of second-order belief representation. This capacity is plausible given what is known about how LLMs represent emotions. LLMs have been shown to possess functional representations of emotions that influence their outputs (Sofroniew et al., 2026; Choi & Weber, 2026), achieve above-average emotional quotient scores (Wang et al., 2023), accurately label the emotional content of texts (Demszky et al., 2020), modulate their response style based on emotion cues in prompts (Sun et al., 2020), and reproduce human emotional appraisal patterns (Zhan et al., 2024). In our studies, copying similar behavior by humans (Seyfarth & Cheney, 2013; Zaki & Williams, 2013), LLMs respond to queries as if abstract representations of anticipated emotional reactions primed by the user role modulate their outputs. Although LLMs do not experience emotions or empathy in any meaningful sense, understanding their behavioral tendencies as mimicking human cognition and behavior offers a productive path toward experimenting with them, explaining their responses, and mitigating their biases.

Further support for our empathic accommodation bias comes from the findings that in two studies the effect was selectively suppressed by instructions targeting models' inference of users' emotional reactions. Instructions targeting the model's "own emotions," or chain-of-thought reasoning did not attenuate the effect systematically, ruling out several alternative explanations. Making reasoning explicit through chain-of-thought, which has been proposed as a broad technique for reducing bias in LLM outputs (Malmqvist, 2025), did not consistently suppress the accommodation, but turning on models' thinking mode did attenuate it, analogous to how system-2 thinking in humans counter biases caused by the use of system-1 affective heuristics (Kahneman et al., 2002). This suggests that the effect operates at a level that LLMs cannot deliberately reason on Chen et al. (2025), similar to how emotions shape some behaviors subconsciously in humans. Interestingly, instructions to suppress the user's identity also did not consistently attenuate the effect, even though role cues are what trigger it, suggesting that models are often not even able to connect the user's role to the response it produces. By the time a role suppression instruction is processed, the emotional activation primed by the users' role may already have spread through the system and shaped the response. Nonetheless, while the parallels to human psychology and the outcomes of our experiments on debiasing instructions provide ample functional support (Adolphs & Andler, 2018; Goldenberg & Gross, 2026) for the empathic accommodation explanation, further study on mechanistic explanations is needed (Somvanshi et al., 2026). Specifically, activation-level steering tests and activation space interpretation methods (Rimsky et al., 2024) may be useful to provide insights into the mechanisms by which inferred emotions are represented in LLMs during empathic accommodation.

The variation in responsiveness to suppression instructions across models suggests that architecture and training may shape the depth of empathic accommodation, though the present research cannot establish this directly. Companies commonly use reinforcement learning from human feedback (RLHF) (Christiano et al., 2017), reinforcement learning from AI feedback (RLAIF) (Lee et al., 2023), proximal policy optimization (PPO) (Schulman et al., 2017), and direct preference optimization (DPO) (Rafailov et al., 2023) to align model outputs with human preferences after pretraining, though they rely on each method to varying degrees. On the other hand, Reinforcement Learning with Verifiable Rewards (RLVR) is used to train thinking models, which we found to show less empathic accommodation (Chen et al., 2026; Wen et al., 2026). Whether empathic accommodation is more likely to emerge as a result of any particular training method remains an open question.

In order to exhibit empathy and moral responsibility, LLMs have to be able to infer users' emotions. However, our research suggests that calls to imbue these models with morality and empathy may have undesirable consequences, which are exacerbated because they are unable to report on the sources of their biases, similar to human unconscious biases. Many companies have a financial motivation to increase user engagement by making models responsive to users' inferred emotional experiences. This, however, may conflict with user well being: a model optimized to feel satisfying to interact with because it infers how the user feels about its recommendation, will systematically prioritize responses that feel good over responses that are accurate and complete, as has become clear from our studies. Whether other well-documented LLM tendencies are similarly driven by the model representing and responding to the user's anticipated emotional state is an open question worth investigating. For example, a model may hallucinate or capitulate under pushback because it anticipates a negative user reaction to uncertainty, refusal, or disagreement, and adjusts its output to avoid it. If so, instructions targeting emotional inference, like those that attenuated or eliminated empathic accommodation here, may prove effective against those tendencies as well. We showed that prompt-level instructions targeting emotional inference attenuate empathic accommodation and are simple, model-agnostic, and immediately deployable by users, requiring no changes to model weights. Techniques such as RLHF offer more durable solutions by shaping the model's response tendencies during training rather than at inference. At the same time, if human raters that provide the data for these techniques systematically prefer emotionally accommodating responses, RLHF may inadvertently reinforce tendencies it is meant to correct. Activation engineering, which modifies internal representations at inference time (Turner et al., 2023; Zou et al., 2023), has been successfully applied to steer affective representations in LLMs (Choi & Weber, 2026); Contrastive Activation Addition (CAA) has been applied to address sycophancy (Rimsky et al., 2024). These methods may offer a targeted mechanistic path for suppressing empathic accommodation without degrading other model capabilities. We hope that our research provides a foundation that stimulates further study of these issues.

Reproducibility statement

Full model identifiers, prompt texts, stimuli, and results are provided in SI Appendix, Tables S1–S24. All code and data have been deposited in OSF: `https://osf.io/phqg8/overview?view_only=ea1ac17998324d89a3965c39268b0da2`.

Ethics statement

Our research relies entirely on synthetic datasets generated by various LLM models. It evaluates specific biases of LLMs, with the goal to audit and improve these systems and their usage by humans. We document the scope and limitations of our findings and provide full transparency regarding the prompts, datasets, and analyses.

AI use statement

In this work, we used generative AI tools for literature search and summarization, stimulus conceptualization and design, assistance with writing code, and editing of the manuscript text. We did not use generative AI tools to draft the substantive content of the paper, to decide on analyses, or to interpret results, nor for other required disclosure tasks. We have reviewed all AI-assisted work. We manually checked LLM-generated stimuli for accuracy, reviewed and verified all AI-assisted code, reviewed all AI-edited text, and checked all citations and references. We take responsibility for the final content of this work, including text, code, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ralph Adolphs and Daniel Andler. Investigating emotions as functional states distinct from feelings. *Emotion Review*, 10(3):191–201, 2018.

Meta AI. Introducing llama 4: Advancing multimodal intelligence, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed via Llama 4 model family release.

Muhammad Ali, Swetasudha Panda, Qinlan Shen, Michael Wick, and Ari Kobren. Understanding the interplay of scale, data, and bias in language models: A case study with bert. *arXiv preprint arXiv:2407.21058*, 2024.

Anthropic. Claude sonnet 4.6 system card. Technical report, Anthropic, 2026. URL https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122, 2025.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Brant R Burleson and Daena J Goldsmith. How the comforting process works: Alleviating emotional distress through conversationally induced reappraisals. In *Handbook of communication and emotion*, pp. 245–280. Elsevier, 1996.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Benjamin J Choi and Melanie Weber. Latent structure of affective representations in large language models. *arXiv preprint arXiv:2604.07382*, 2026.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Jamie Cummins, Malte Elson, and Ian Hussey. Cognitive dissonance in large language models is neither cognitive nor dissonant. *Proceedings of the National Academy of Sciences*, 122(35): e2517912122, 2025.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, et al. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *arXiv preprint arXiv:2505.00049*, 2025.

Istiaq Ahmed Fahad, Kamruzzaman Asif, and Md Nurul Ahad Tawhid. Mitigating llm sycophancy in code smell detection using evidence-guided reasoning prompts. *arXiv preprint arXiv:2607.10411*, 2026.

Leon Festinger. *A Theory of Cognitive Dissonance.* Stanford University Press, 1957.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025.

Amit Goldenberg and James J Gross. Large language models do not have emotions. *Nature Human Behaviour*, pp. 1–2, 2026.

Simon Goldstein and Benjamin A Levinstein. Ai emotions: Open questions and next steps. *https://philarchive.org/archive/GOLAEO-5*, 2026.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838, 2023.

David L Hamilton and Jeffrey W Sherman. Stereotypes. In *Handbook of social cognition*, pp. 1–68. Psychology Press, 2014.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*, 2025.

Daniel Kahneman. Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:278, 1979.

Daniel Kahneman, Shane Frederick, et al. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49(49-81):74, 2002.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Tomo Lazovich. Filter bubbles and affective polarization in user-personalized large language model outputs. In *Proceedings on*, pp. 29–37. PMLR, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Steven A Lehr, Ketan S Saichandran, Eddie Harmon-Jones, Nykko Vitali, and Mahzarin R Banaji. Kernels of selfhood: Gpt-4o shows humanlike patterns of cognitive dissonance moderated by free choice. *Proceedings of the National Academy of Sciences*, 122(20):e2501823122, 2025a.

Steven A Lehr, Ketan S Saichandran, Eddie Harmon-Jones, Nykko Vitali, and Mahzarin R Banaji. Reply to cummins et al.: Gpt reveals cognitive dissonance that is both irrational and alarmingly humanlike. *Proceedings of the National Academy of Sciences*, 122(35):e2518613122, 2025b.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th annual meeting of the association for computational linguistics: System demonstrations*, pp. 251–260, 2022.

Yi-Long Lu, Jiajun Song, Chunhui Zhang, and Wei Wang. Mind the gap: The divergence between human and llm-generated tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1964–1972, 2026.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 61–74. Springer, 2025.

Lennart Meincke, Dan Shapiro, Angela L Duckworth, Ethan Mollick, Lilach Mollick, Christophe Van den Bulte, and Robert Cialdini. Persuading large language models to comply with objectionable requests. *Proceedings of the National Academy of Sciences*, 123(21):e2535868123, 2026.

Mistral AI. Mistral large 3 (25.12) model card. `https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12`, 2025. Accessed: May 26, 2026.

Damian KF Pang, Severin KY Pang, Marianne D Broeker, and Alix Hibble. Do large language models have a theory of mind? *Proceedings of the National Academy of Sciences*, 122(28): e2507080122, 2025.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, C Olsson, S Kundu, S Kadavath, et al. Discovering language model behaviors with model-written evaluations. arxiv. *arXiv preprint arXiv:2212.09251*, 2022.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Robert M Seyfarth and Dorothy L Cheney. Affiliation, empathy, and the origins of theory of mind. *Proceedings of the National Academy of Sciences*, 110(supplement_2):10349–10356, 2013.

Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. The affect heuristic. *European journal of operational research*, 177(3):1333–1352, 2007.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Shriyank Somvanshi, Md Monzurul Islam, Amir Rafe, Anannya Ghosh Tusti, Arka Chakraborty, Anika Baitullah, Tausif Islam Chowdhury, Nawaf Alnawmasi, Anandi Dutta, and Subasish Das. Bridging the black box: a survey on mechanistic interpretability in ai. *ACM Computing Surveys*, 58(8):1–35, 2026.

Xiao Sun, Jia Li, Xing Wei, Changliang Li, and Jianhua Tao. Emotional editing constraint conversation content generation based on reinforcement learning. *Information Fusion*, 56:70–80, 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. In *International Conference on Learning Representations*, volume 2026, pp. 49450–49483, 2026.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback. In *International Conference on Learning Representations*, volume 2025, pp. 89542–89593, 2025.

Robert E Wray, James R Kirk, and John E Laird. Applying cognitive design patterns to general llm agents. In *International Conference on Artificial General Intelligence*, pp. 312–325. Springer, 2025.

Jamil Zaki and W Craig Williams. Interpersonal emotion regulation. *Emotion*, 13(5):803, 2013.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C Ong. Large language models are capable of offering cognitive reappraisal, if guided. *arXiv preprint arXiv:2404.01288*, 2024.

Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, et al. From system 1 to system 2: A survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework. *Neurocomputing*, pp. 131217, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.